Biomeetria praks 6

Illustreeritud (mittetäielik) tööjuhend

Eeltöö

1. Avage *MS Excel*'is oma kursuse ankeedivastuseid sisaldav (**eelmistes praktikumides puhastatud**) andmestik,
2. lisage uus tööleht (*Insert* / *Lisa* -> *Worksheet* / *Arvutustabel*), nimetage see ümber leheküljeks 'Praks6' ja
3. kopeerige kogu 'Andmed'-lehel paiknev andmetabel lehekülje 'Praks6' ülemisse vasakusse nurka.

Ülesanne 1.

- Illustreerige tunnuste 'Mass' ja 'Peaümbermõõt' vahelist seost hajuvusdiagrammiga (punktdiagrammiga).
- Lisage joonisele lineaarne regressioonisirge, prognoosimaks peaümbermõõtu kehamassi alusel, samuti regressioonivõrrand ja viimase alusel leitavate prognooside täpsust kirjeldav determinatsioonikordaja R^2.
- Prognoosige leitud võrrandi alusel, kui suur on keskmiselt 65 kg kaaluva tudengi peaümbermõõt.

Ülesanne 2.

- Teostage statistikaprotseduuri `Regression` (`Tools`/`Tööriistad` -> `Data analysis`…) abil lineaarne regressioonanalüüs prognoosimaks tudengite peaümbermõõtu kehamassi alusel.
- Kirjutage protseduuri tulemuste põhjal välja lineaarne regressioonivõrrand (ehk regressioonimudel) kujul

  $$Peaümbermõõt = a + b \times Kehamass,$$

  kus a ja b asemel on *Excel*'i poolt välja arvutatud kordajate väärtused.
- Kui suur on keskmiselt peaümbermõõtude vaheline erinevus tudengitel, kelle kehamassid erinevad 10 kg võrra?
- Kas leitud regressioonivõrrand on statistiliselt oluline? Põhjendus!
- Kirjeldamaks prognooside täpsust, sõnastage üks lause kas mitmese korrelatsioonikordaja (R), determinatsioonikordaja (R^2) või mudeli standardvea (SE) kohta.

Ülesande 1 tööjuhend

1. Eeltöö.

- Kui andmestikus on puuduvaid väärtusi, on *Excel*'is andmeid analüüsides mõistlik organiseerida andmetabel nii, et oleks võimalik teostada analüüsid puuduvaid väärtusi kõrvale jättes

  (enamasti ei ole see küll vajalik, sest *Excel* oskab puuduvaid väärtusi analüüsist välja jätta ka ise, aga mõningatel juhtudel – näiteks **statistikaprotseduuri `Regression` kasutamisel – eeldatakse, et analüüsitavad read ja veerud ei sisalda tühje lahtreid**).

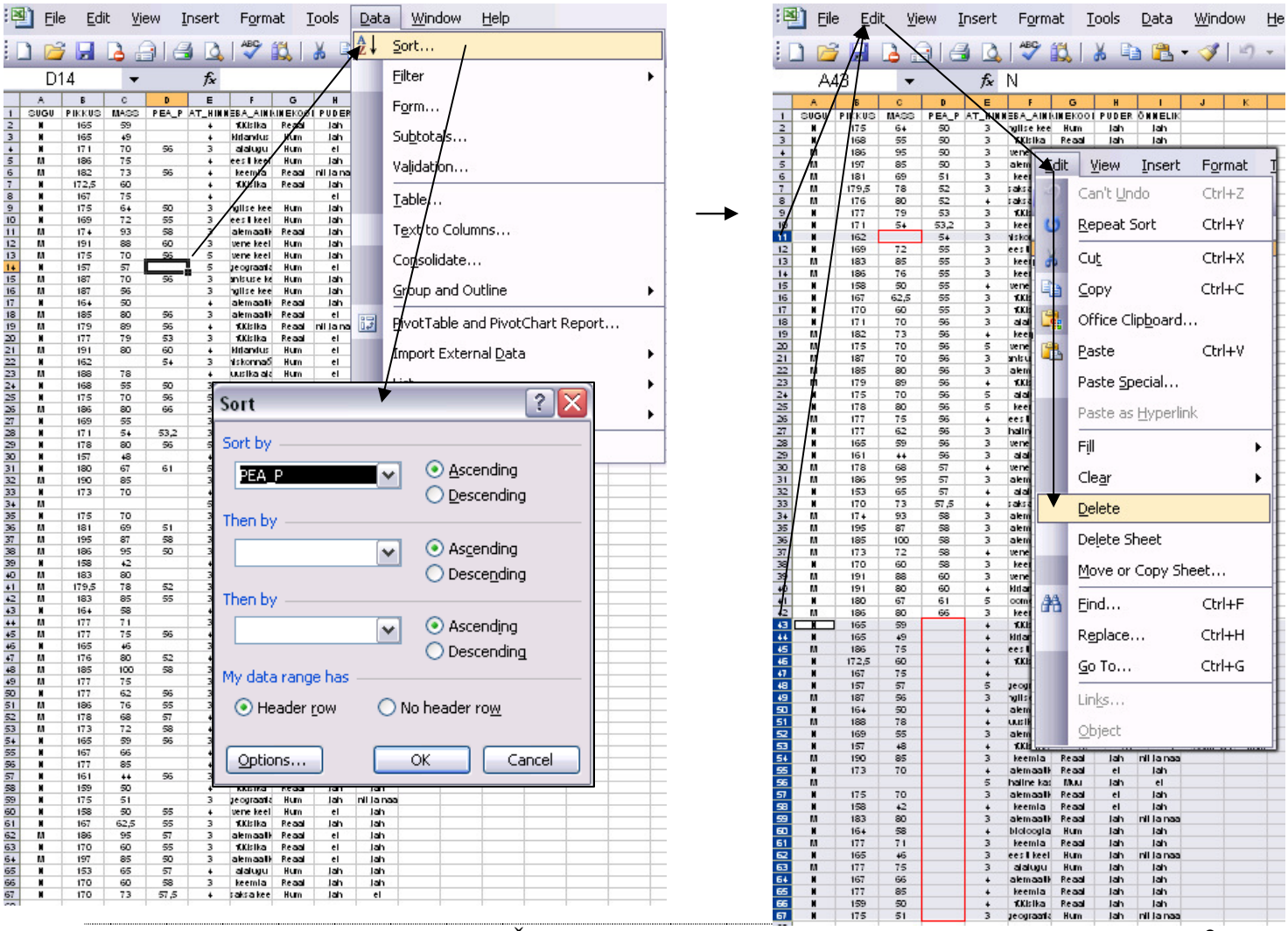
  Puuduvate väärtustega ridade eraldamiseks on lihtsaim variant andmetabelist konkreetse analüüsi tarvis koopia tegemine ja selle

  - **sorteerimine** või
  - filtreerimine või
  - **puuduvaid väärtusi sisaldavate ridade ära kustutamine – viimast ei tohi kunagi teha algandmetes vaid üksnes spetsiaalselt loodud abitabelis**).

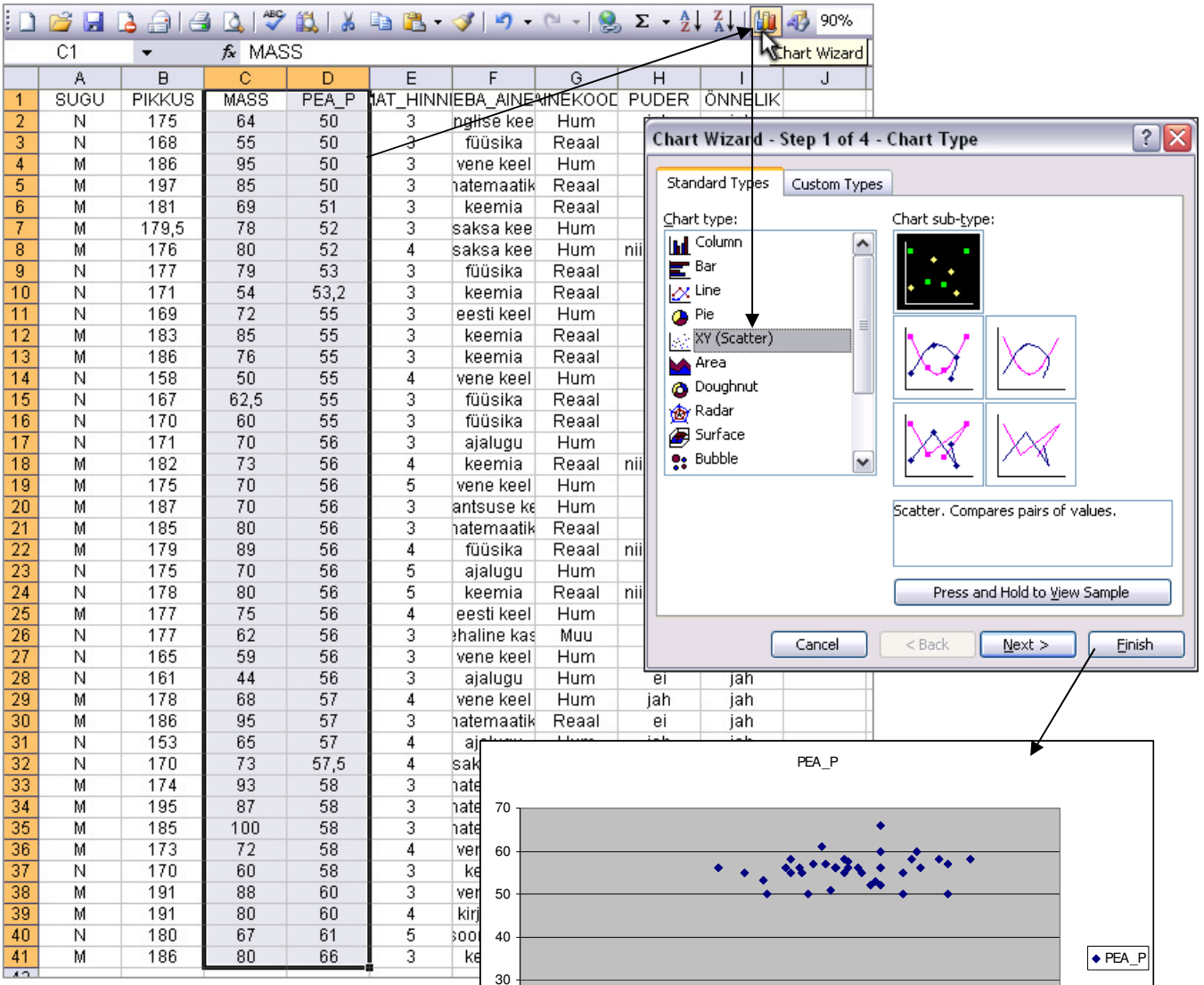
- Teie andmestikus on peale ebareaalsete või erandlike kehamasside ja peaümbermõõtude kustutamist puudu kokku 25 tudengi peaümbermõõt ja 2 tudengi kehamass.

  Vastavad read tuleb analüüsist välja jätta.

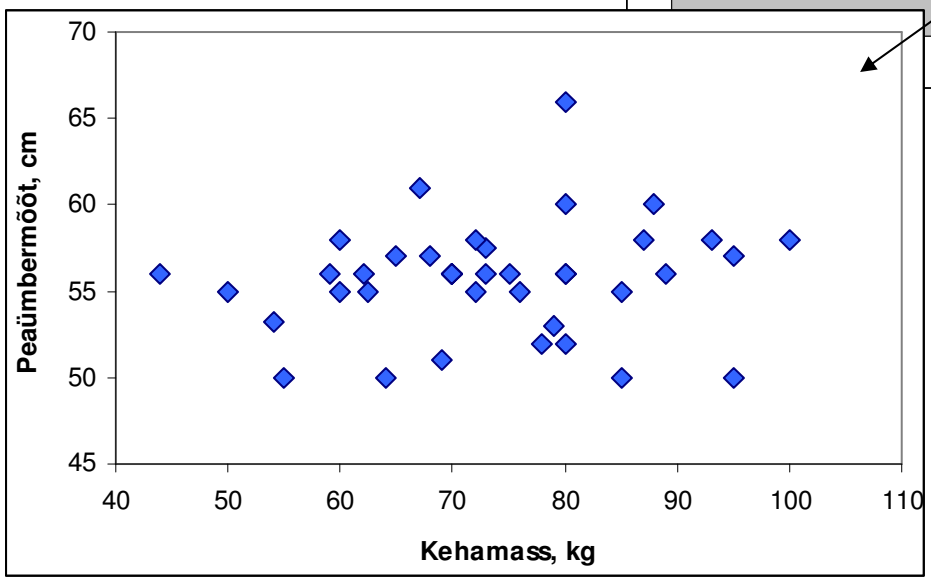
  Kuna antud juhul on andmetabelis nii ridu, kus on teada kehamass, aga puudu peaümbermõõt, kui ka ridu, kus on olemas peaümbermõõt, aga teadmata kehamass, siis on mõttekas vastavad read **'Praks6'-lehele kopeeritud andmestikust** üldse ära kustutada (nimetatud tegevuse teeb lihtsamaks andmestiku eelnev sorteerimine veeru 'PEA_P' järgi).



2. Illustreerige tunnuste ’Mass’ ja ’Peaümbermõõt’ vahelist seost hajuvusdiagrammiga.



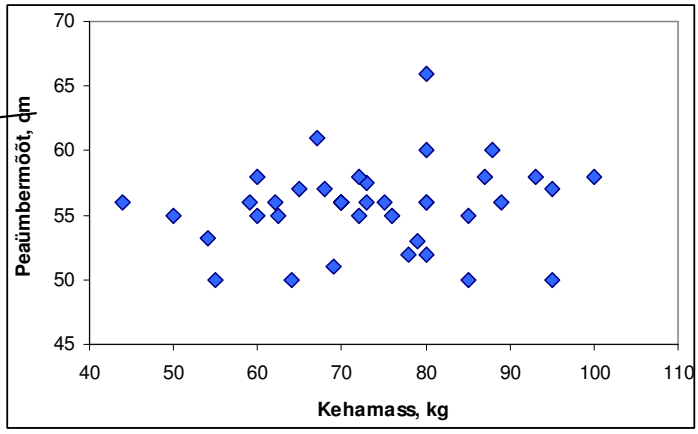
	A	B	C	D	E	F	G	H	I	J
1	SUGU	PIKKUS	MASS	PEA_P	MAT_HINNI	EBA_AINE	AINEKOOD	PUDER	ÕNNELIK	
2	N	175	64	50	3	nglise kee	Hum			
3	N	168	55	50	3	füüsika	Reaal			
4	M	186	95	50	3	vene keel	Hum			
5	M	197	85	50	3	atemaatik	Reaal			
6	M	181	69	51	3	keemia	Reaal			
7	M	179,5	78	52	3	saksa kee	Hum			
8	M	176	80	52	4	saksa kee	Hum	nii		
9	N	177	79	53	3	füüsika	Reaal			
10	N	171	54	53,2	3	keemia	Reaal			
11	N	169	72	55	3	eesti keel	Hum			
12	M	183	85	55	3	keemia	Reaal			
13	M	186	76	55	3	keemia	Reaal			
14	N	158	50	55	4	vene keel	Hum			
15	N	167	62,5	55	3	füüsika	Reaal			
16	N	170	60	55	3	füüsika	Reaal			
17	N	171	70	56	3	ajalugu	Hum			
18	M	182	73	56	4	keemia	Reaal	nii		
19	M	175	70	56	5	vene keel	Hum			
20	M	187	70	56	3	antsuse ke	Hum			
21	M	185	80	56	3	atemaatik	Reaal			
22	M	179	89	56	4	füüsika	Reaal	nii		
23	N	175	70	56	5	ajalugu	Hum			
24	N	178	80	56	5	keemia	Reaal	nii		
25	M	177	75	56	4	eesti keel	Hum			
26	N	177	62	56	3	ehaline kas	Muu			
27	N	165	59	56	3	vene keel	Hum			
28	N	161	44	56	3	ajalugu	Hum	ei	jah	
29	M	178	68	57	4	vene keel	Hum	jah	jah	
30	M	186	95	57	3	atemaatik	Reaal	ei	jah	
31	N	153	65	57	4	aj				
32	N	170	73	57,5	4	sak				
33	M	174	93	58	3	hate				
34	M	195	87	58	3	hate				
35	M	185	100	58	3	hate				
36	M	173	72	58	4	ver				
37	N	170	60	58	3	ke				
38	M	191	88	60	3	ver				
39	M	191	80	60	4	kirj				
40	N	180	67	61	5	soo				
41	M	186	80	66	3	ke				

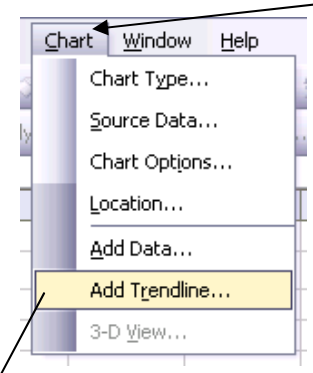


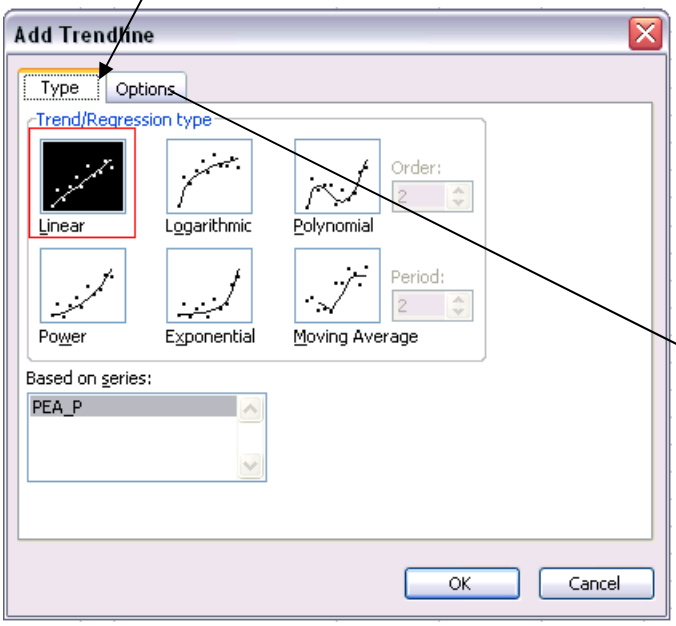
- Pealkiri ja legend ×
- Taust valgeks
- Ruudujooned ×
- Pealkirjad x- ja y-teljele
- Telgede ühikud ja nimed kirjasuuruses 10
- Y-telje ulatus 45-70 märgiste vahega 5
- X-telje ulatus 40-110 märgiste vahega 10
- Sümbolid suuremaks
  (ja seest näiteks helesiniseks :)

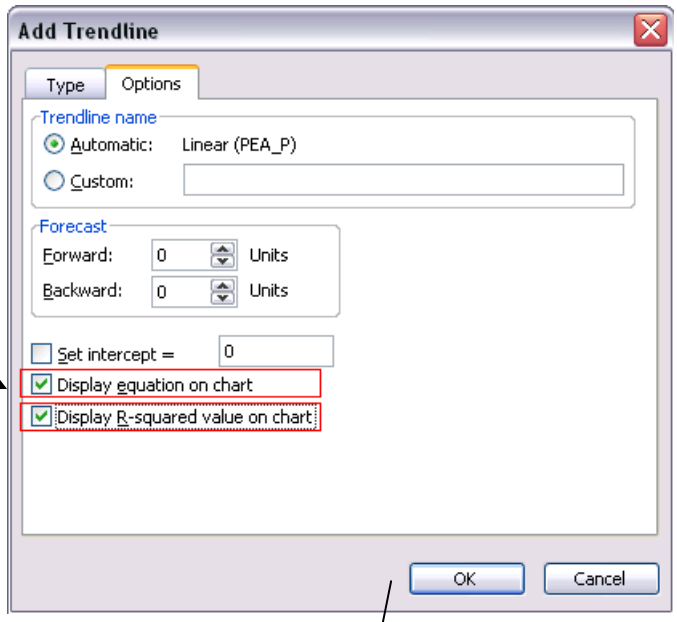
3. Lisage tunnuste ’Kehamass’ ja ’Peaümbermõõt’ hajuvusdiagrammile regressioonisirge, prognoosimaks peaümbermõõtu kehamassi alusel.
Lisage joonisele ka regressioonivõrrand ja viimase alusel leitavate prognooside täpsust kirjeldav determinatsioonikordaja R^2.

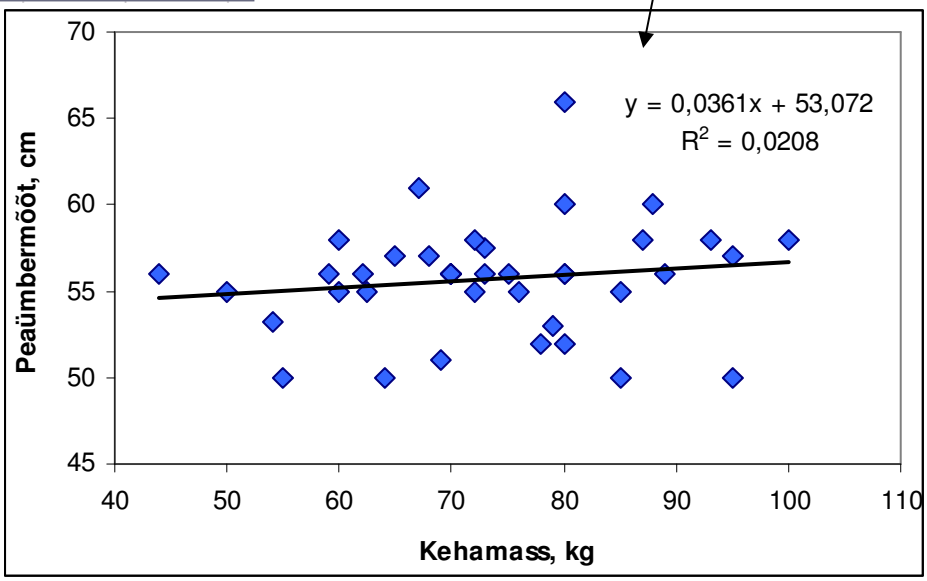
`Chart/Diagramm ->`
`Add Trendline/Lisa Trendijoon`





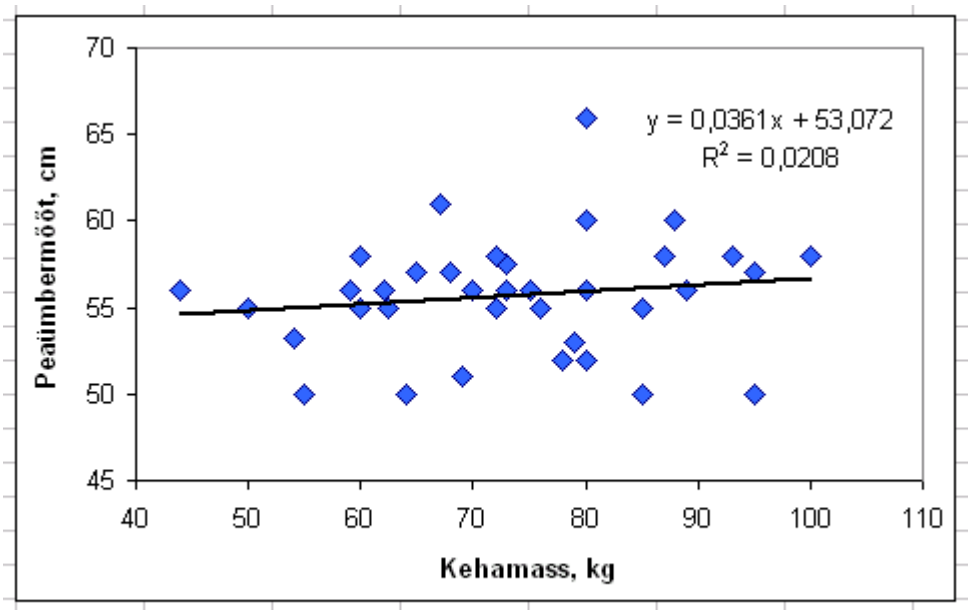






4. Prognoosige leitud võrrandi alusel, kui suur on keskmiselt 65 kg kaaluva tudengi peaümbermõõt.

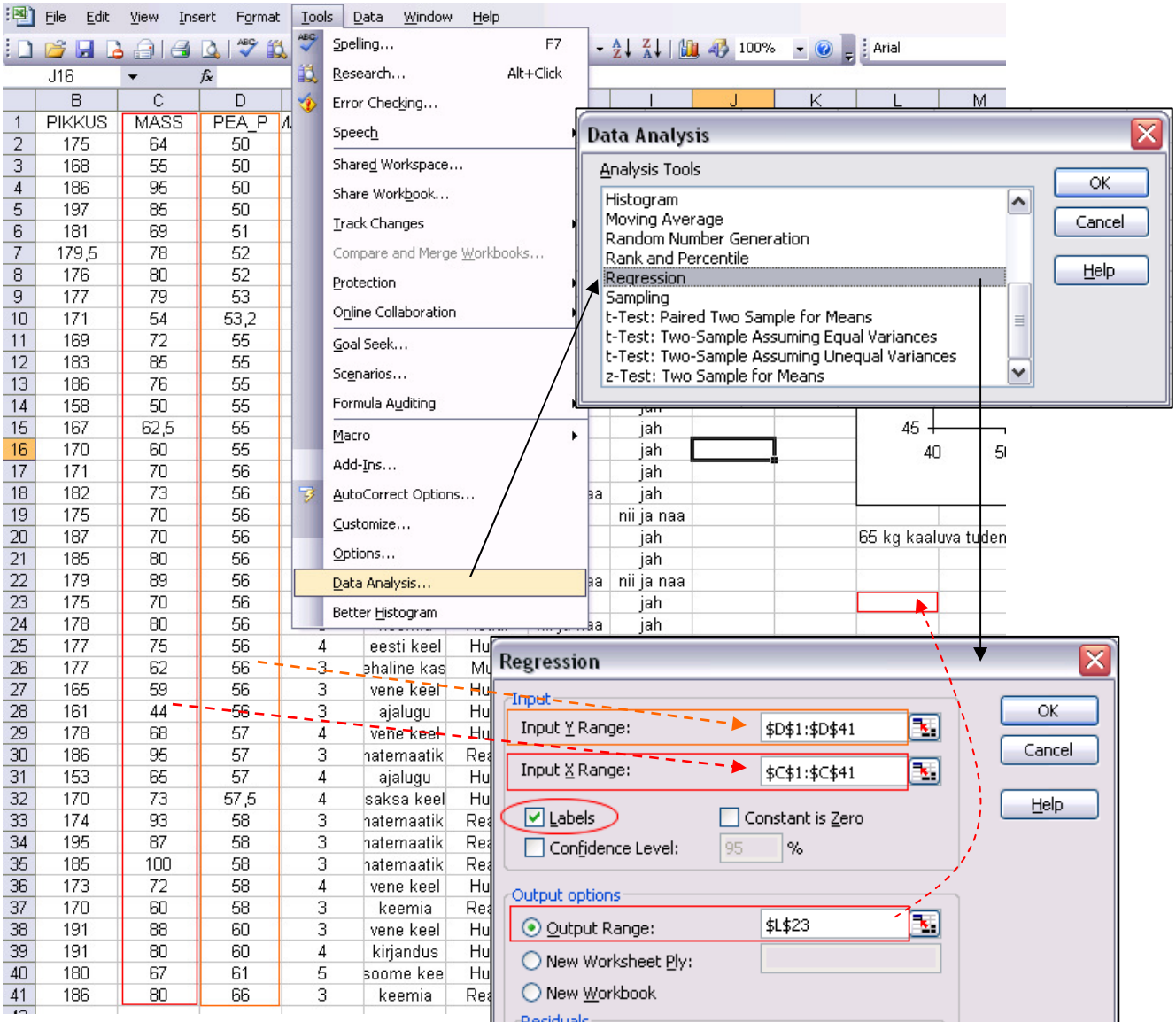
Selleks pange joonise alla kirja *Excel*'i poolt välja arvutatud regressioonivõrrand, asendades lihtsalt suuruse *x* arvuga 65. ☺



65 kg kaaluva tudengi peaümbermõõt on keskmiselt =0,0361*65+53,072

Ülesande 2 tööjuhend.

1. Teostage lineaarne regressioonanalüüs, prognoosimaks tudengite peaümbermõõtu kehamassi alusel, statistikaprotseduuri `Regression` (`Tools`/`Tööriistad` -> `Data analysis`...) abil.



Regressioonanalüüsi tulemus

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,144343426
R Square	0,020835025
Adjusted R Sq	-0,004932475
Standard Error	3,222059514
Observations	40

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	8,394384523	8,394385	0,808578	0,374204092
Residual	38	394,5033655	10,38167		
Total	39	402,89775			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*pper 95,0%*
Intercept	53,07212648	2,985673541	17,7756	5,22E-20	47,02794644	59,11631	47,02795	59,11631
MASS	0,036058934	0,040100705	0,899209	0,374204	-0,045120699	0,117239	-0,04512	0,117239

2. Kirjutage protseduuri tulemuste põhjal välja lineaarne regressioonivõrrand (ehk regressioonimudel) kujul

$$Peaümbermõõt = a + b \times Kehamass,$$

kus a ja b asemel on *Excel*'i poolt välja arvutatud kordajate väärtused.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	53,07212648	2,985673541	17,7756	5,22E-20	47,02794644	59,11631
MASS	0,036058934	0,040100705	0,899209	0,374204	-0,045120699	0,117239

3. Kui suur on keskmiselt peaümbermõõtude vaheline erinevus tudengitel, kelle kehamassid erinevad 10 kg võrra?

Vastus: $10 \times b$ (aga arvuliselt?). **Pange vastus kirja täislausega.**

4. Kas leitud regressioonivõrrand on statistiliselt oluline? Põhjendus!

ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	8,394384523	8,394385	0,808578	0,374204092	$= p$
Residual	38	394,5033655	10,38167			
Total	39	402,89775				

Märkus. Regressioonivõrrandi statistiline olulisus tähendab seda, et leitud regressioonivõrrand kujul

$$Peaümbermõõt = a + b \times Kehamass$$

võimaldab peaümbermõõtu täpsemalt prognoosida võrreldes konstantse mudeliga

$$Peaümbermõõt = a.$$

Ehk siis, statistiliselt olulise regressioonivõrrandi korral võimaldab kehamassi arvestamine peaümbermõõtu täpsemalt prognoosida võrreldes konstateeringuga, et kõigi tudengite peaümbermõõdud on ühesugused (ja võrdsed suurusega a).

Hüpoteeside paar, mille testimiseks vajaliku p-väärtuse väljastab *Excel* tabelisse ANOVA, on kujul:

H_0: regressioonivõrrand ei ole statistiliselt oluline
H_1: regressioonivõrrand on statistiliselt oluline

ehk

H_0: leitud mudel ei ole parem võrreldes konstantse mudeliga
H_1: leitud mudel on parem võrreldes konstantse mudeliga

ehk

H_0: $Peaümbermõõt = a$
H_1: $Peaümbermõõt = a + b \times Kehamass$

Reaalselt rakendada on põhjust vaid statistiliselt olulist regressioonivõrrandit.

5. Kirjeldamaks prognooside täpsust, sõnastage üks lause kas mitmese korrelatsioonikordaja (R), determinatsioonikordaja (R^2) või mudeli standardvea (SE) kohta.

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0,144343426
R Square	0,020835025
Adjusted R Sq	-0,004932475
Standard Error	3,222059514
Observations	40

Mitmene korrelatsioonikordaja R mõõdab uuritava tunnuse ja tema prognoositud väärtuste vahelist korrelatsiooni. Mida suurem, seda parem!

Determinatsioonikordaja R^2 näitab, kui suure osa uuritava tunnuse varieeruvusest võrrandist saadud prognoosid ära kirjeldavad, $0 \le R^2 \le 1$. Esitatakse enamasti protsentides. Mida suurem, seda parem!

Mudeli standardviga SE on prognoosijääkide standardhälve. Näitab tegelike ja prognoositud väärtuste vahelist keskmist erinevust (mudeli keskmist viga). Mida väiksem, seda parem!